## The RAP1 Binding Site Can Be Determined from IP Data Alone

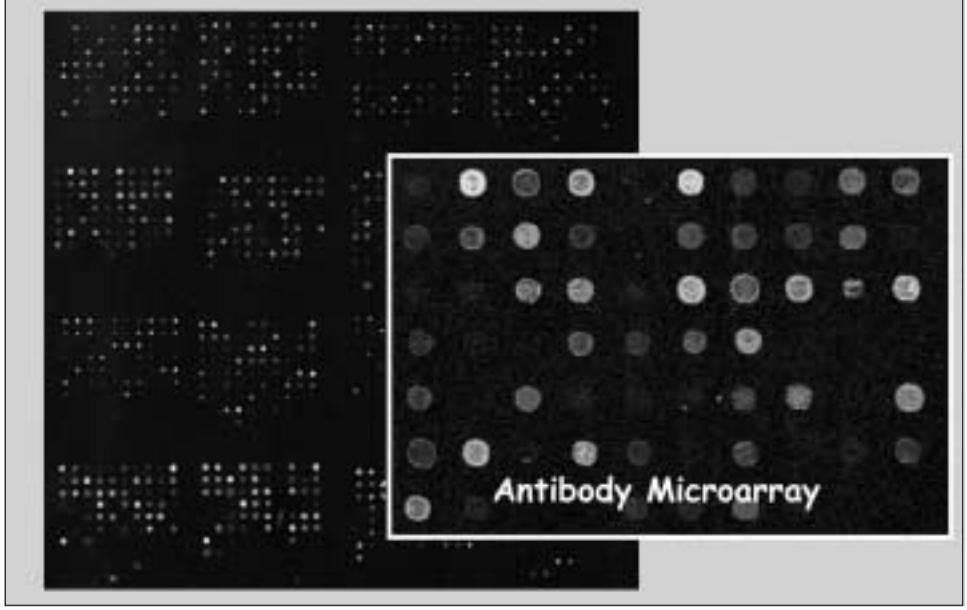
### Published RAP1 Consensus Sites

| | | |
|---|---|---|
| #1 | RTRCACCCANNCMCC | Oligo Selection/Amplification |
| #2 | WACAYCCRTACATY | In Silico w/ Ribosomal Promoters |
| #3 | RMAYCCRMNCAYY | Monomer Binding in Vitro |
| #4 | RMACCCANNCAYY | Original In Vitro/Sequence Data |
| #5 | ACACCCAYACAYYY | Structure/Function Analysis |
| #6 | ACAYYY//ACAYYY | Structure/Function Analysis |

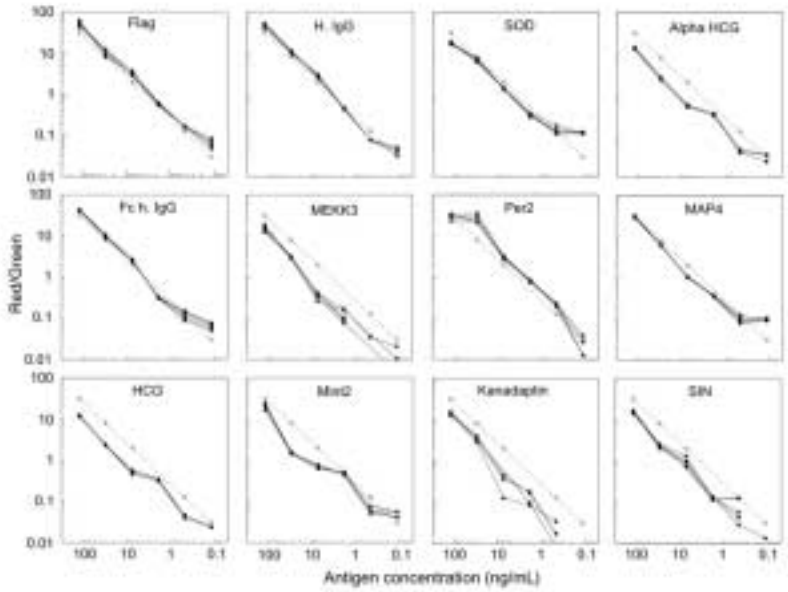
**RCACCCANNCAYY  Overall Consensus**

### Rap1p Binding Sites Determined by IP Data Alone

| | Selected Intergenics only | No Telomerics |
|---|---|---|
| Consensus: | ACACCCATACATC | ACACCCATACATC |
| Degenerate: | ACACCCRTACAYY | ACAYCCRTACAYY |



## Antibody Microarray Performance




## The Dynamic Genome

Every individual of every species has, to a first approximation, only one genome, and with this essentially static genome:

Unicellular organisms can survive, grow and reproduce in rapidly and extremely variable environmental conditions.

Multicellular organisms are capable of producing cells and tissues with dramatically different properties.

# Evolution of Gene Regulation

Evolution ensures that coding genes make proteins with the proper molecular properties: enzymatic activity, binding, etc….

Equally importantly, natural selection has also proceeded to ensure that these proteins are made and function when, where and in the proper form and amounts required, and that they are not made when their presence would be deleterious, or in unnecessary amounts that would waste cellular energy.

## Evolutionary Logic Links Gene Expression and Gene Function

The evolutionary logic that dictates that genes be made only when and where needed implies a direct connection between a gene's pattern of expression and its function

## Evolutionary Logic Links Gene Expression and Cellular State/Phenotype

The important role of gene regulation in development and cellular responses means that the particular set of genes a cell or collection of cells is expressing at any moment can tell us a great deal about its history, environment, internal state and future.

**Gene expression provides a window onto two properties of biological systems that we are extremely interested in studying**

**Gene Function**

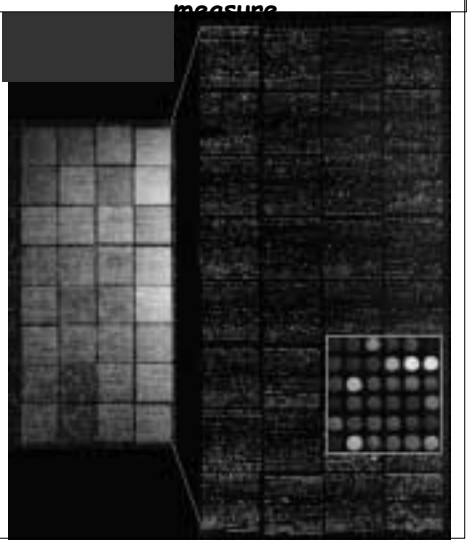
↕

**Gene Expression**

↕

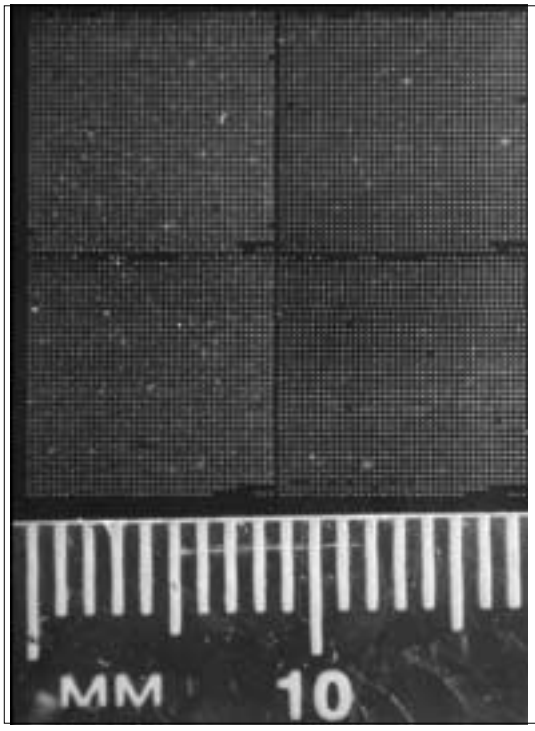
**State/Phenotype**

**Unlike gene function or the physiological state of cells, tissues and organisms, gene expression is now easy to measure**



# Saccharomyces cerevisiae

- Unicellular fungus with extensive commercial importance
- ~12 M genome; sequencing completed 1996
- ~6,200 ORFs
- Despite extensive genetic, molecular and biochemical analysis, when genome sequence completed only ~35% of genes had reasonable functional annotation

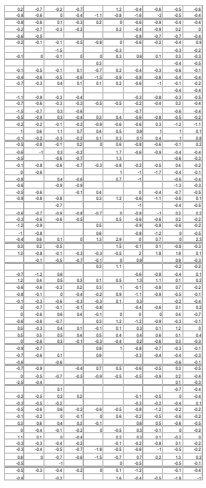



**DNA microarrays with elements representing every identified open reading frame – either in the form of PCR amplified ORFs or synthetic oligonucleotides - have been available for 5 years and are in fairly wide use**

## The Principle Challenge in Experimental Genomics is Making Biological Sense of the Data

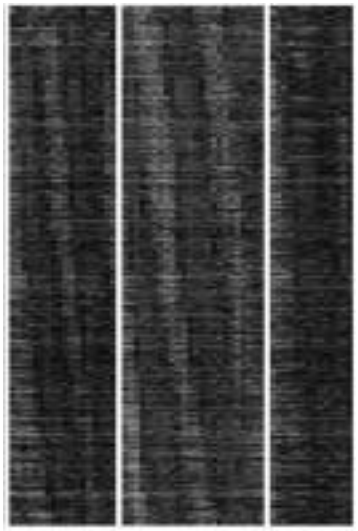
Experiments

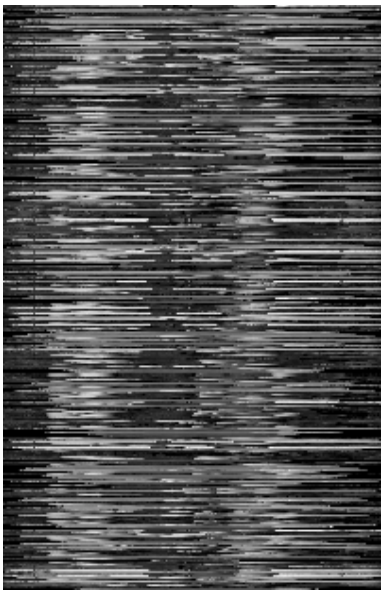
Genes



***Saccharomyces cerevisiae***
**whole genome ORF arrays**

**~6200 genes**
**~6000 conditions**
**> 36 million observations**

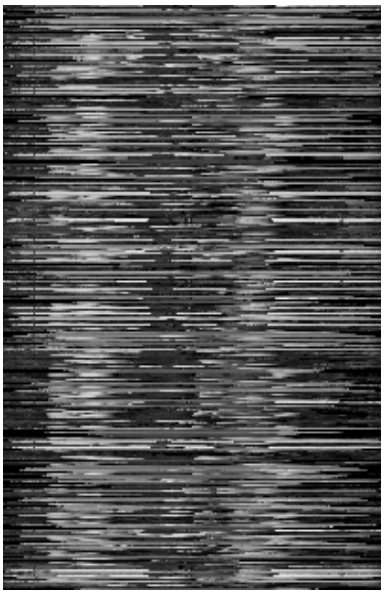


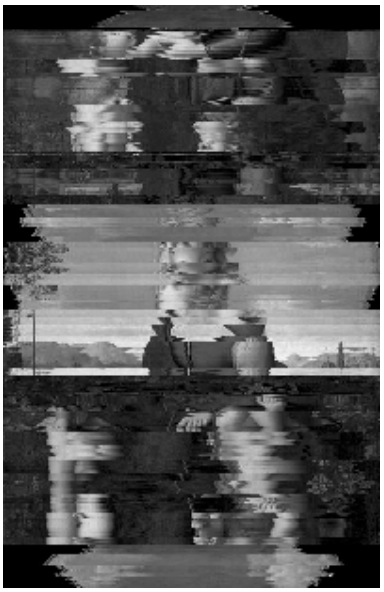


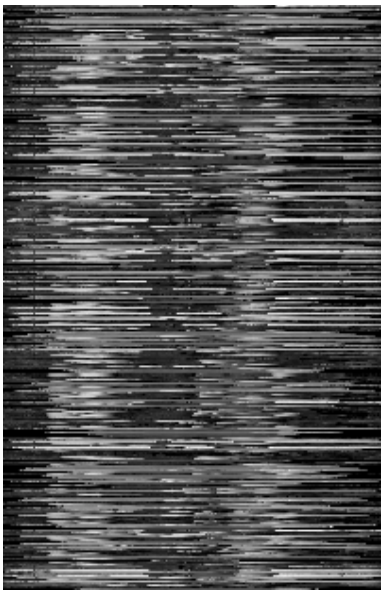
## How Do We Make Biological Sense of Complex Gene Expression Datasets

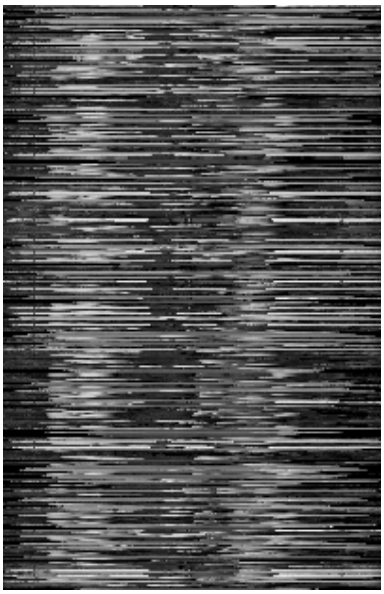
A corollary of this logical relationship between gene expression and gene function is that genes with similar function should have similar patterns of gene expression.

To the extent that this is true, this property can be used to impart a logical and biologically meaningful order to complex gene expression data

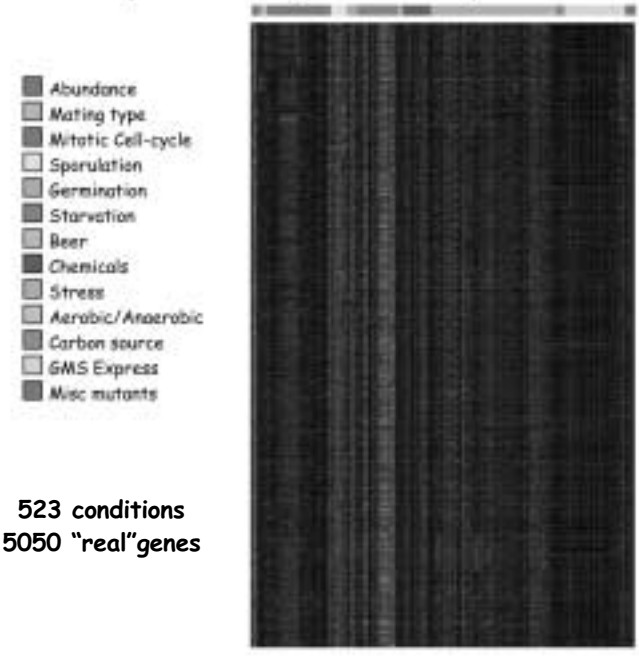






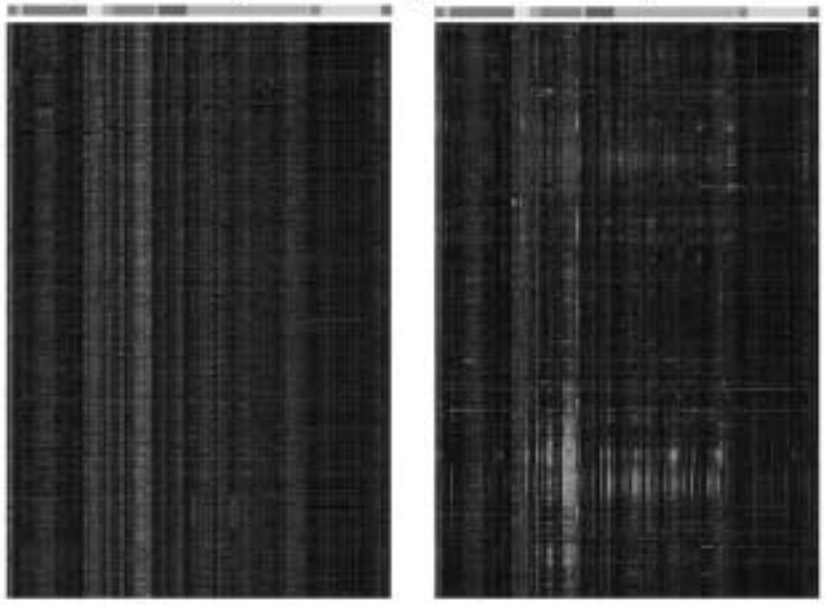


## Gene Expression Program of the yeast Saccharomyces cerevisiae



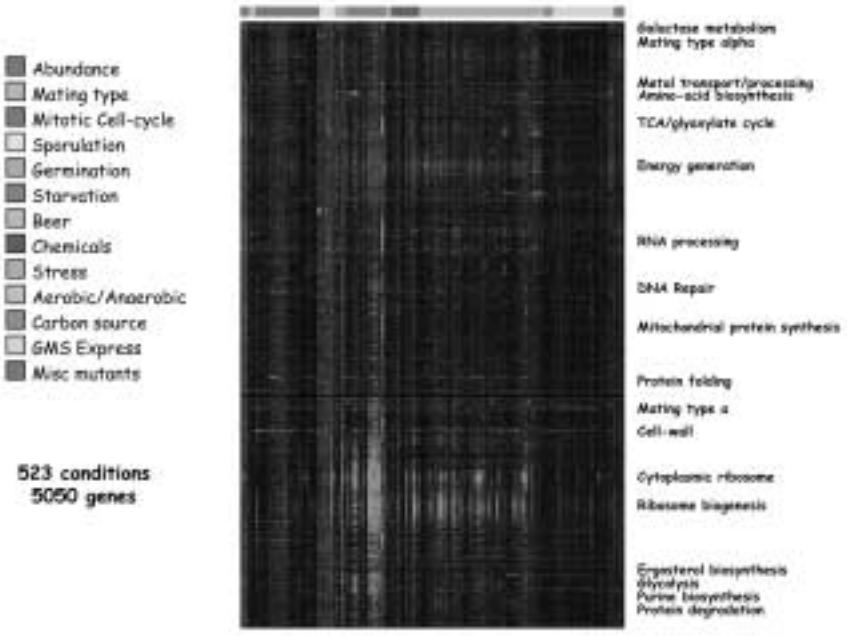

**523 conditions**
**5050 "real"genes**

## Gene Expression Program of the yeast Saccharomyces cerevisiae



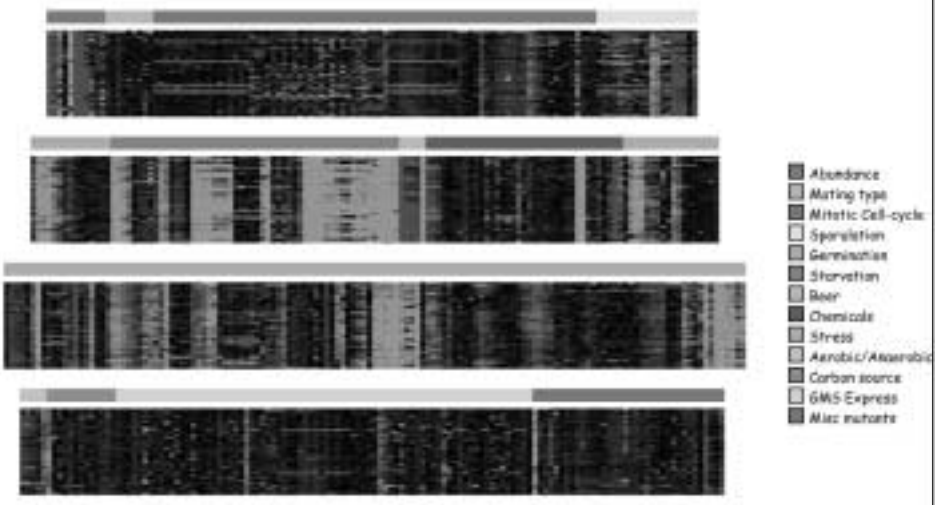
UNCLUSTERED CLUSTERED

## Gene Expression Program of the yeast Saccharomyces cerevisiae



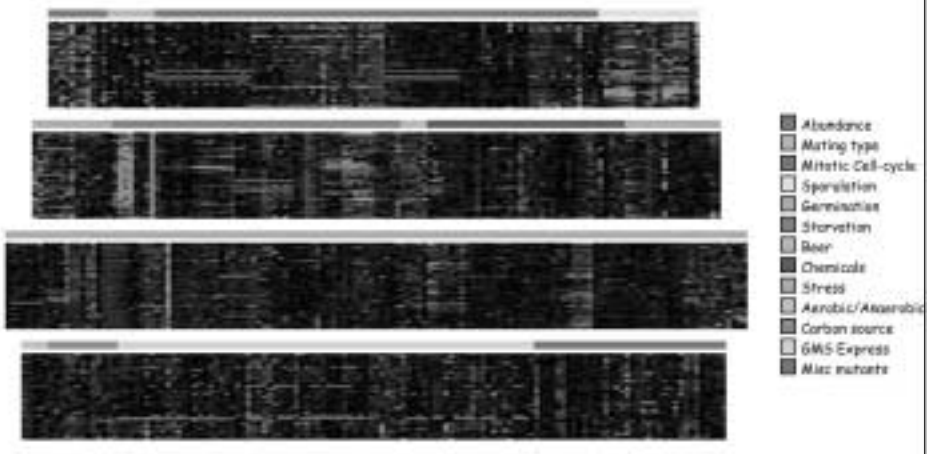
# Cytoplasmic Ribosome Cluster

Cluster contains virtually all subunits of cytoplasmic ribosome



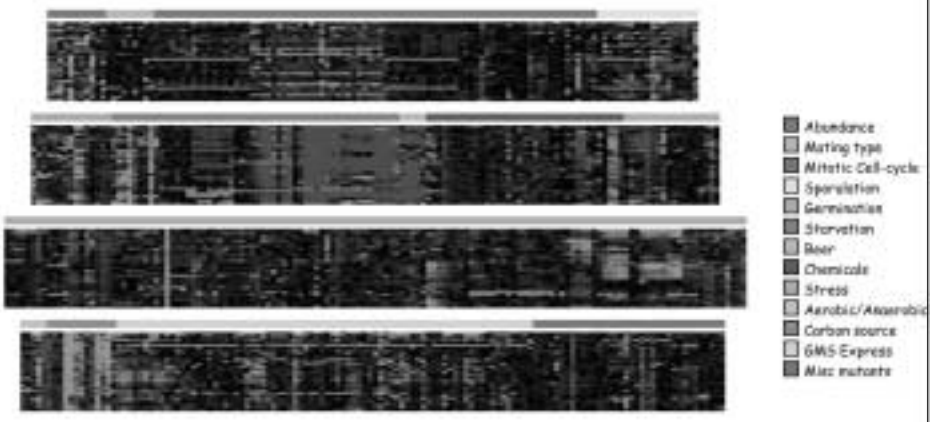

# Mitochondrial Ribosome Cluster

Contains most known mitochondrial ribosome components
plus 8 proteins with no annotated function,
3 with homology to bacterial ribosomal proteins



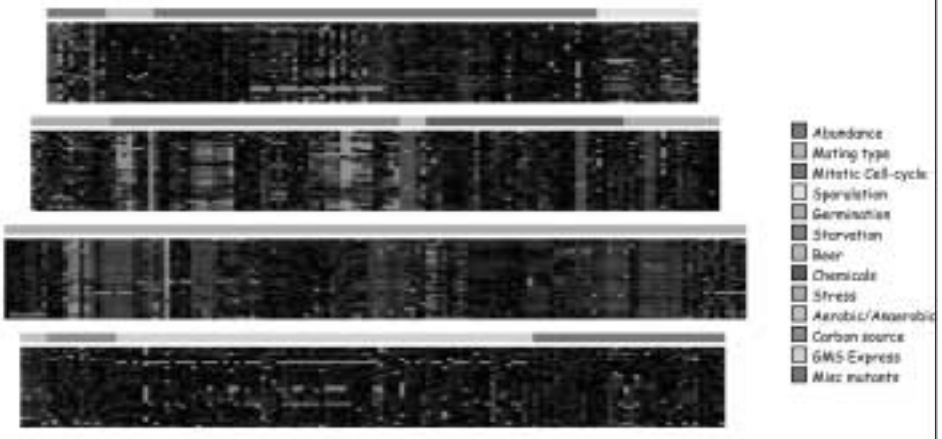
# Energy Generation Cluster

Oxidative Phosphorylation and ATP Synthesis

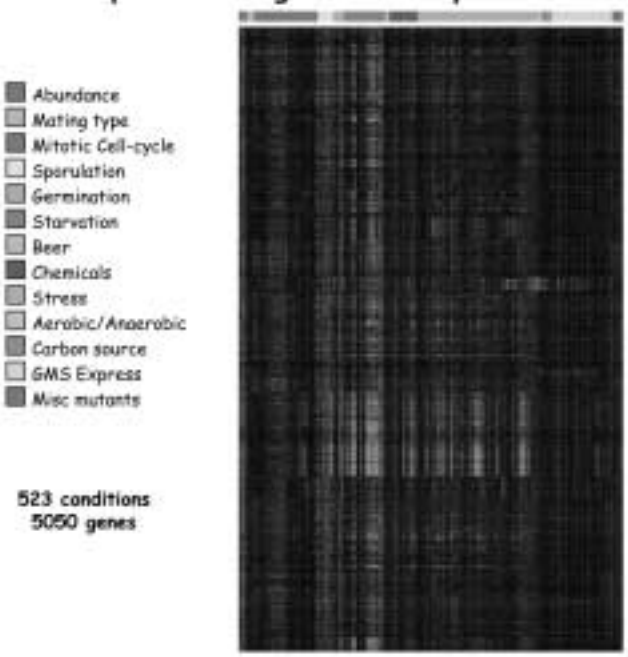



# Proteasome Cluster

Contains virtually all known components of proteasome
(20s and 26s subunits)

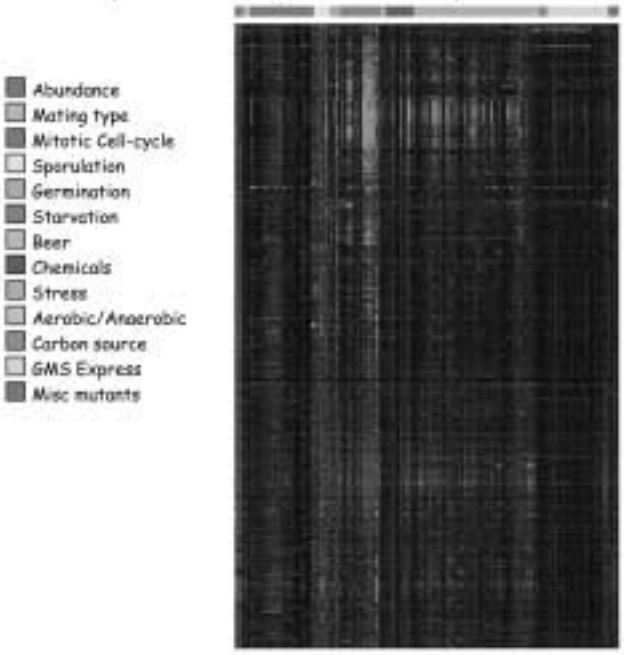


## Gene Expression Program of the yeast Saccharomyces cerevisiae



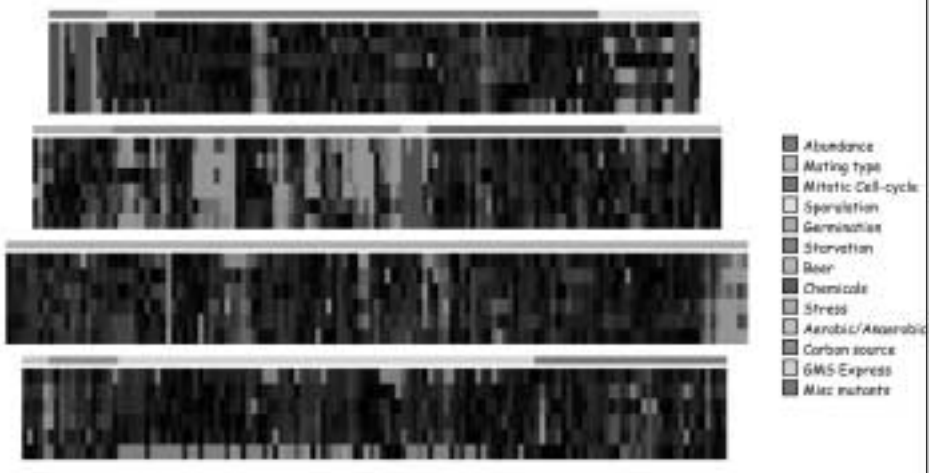

**All known multisubunit complexes**

## Gene Expression Program of the yeast Saccharomyces cerevisiae



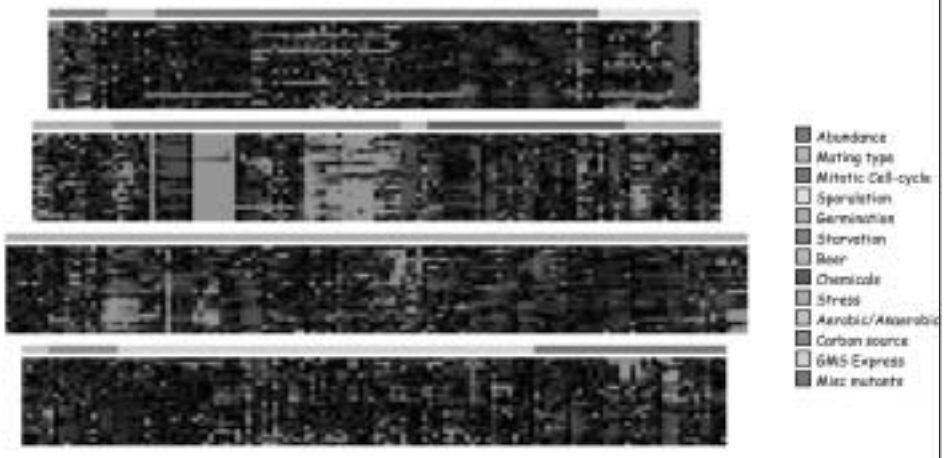

**Proteins not known to be part of multisubunit complexes**

# Glycolysis Cluster



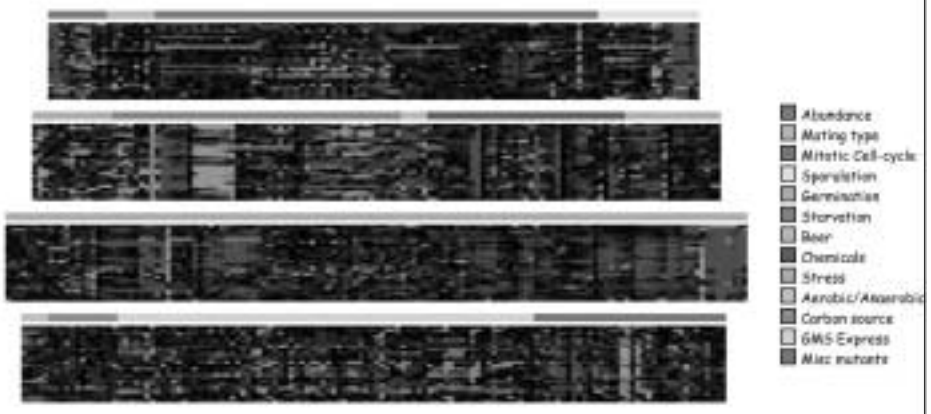

# Amino Acid and Purine Biosynthesis

Lue, Ile, Val, Lys
Purine, His



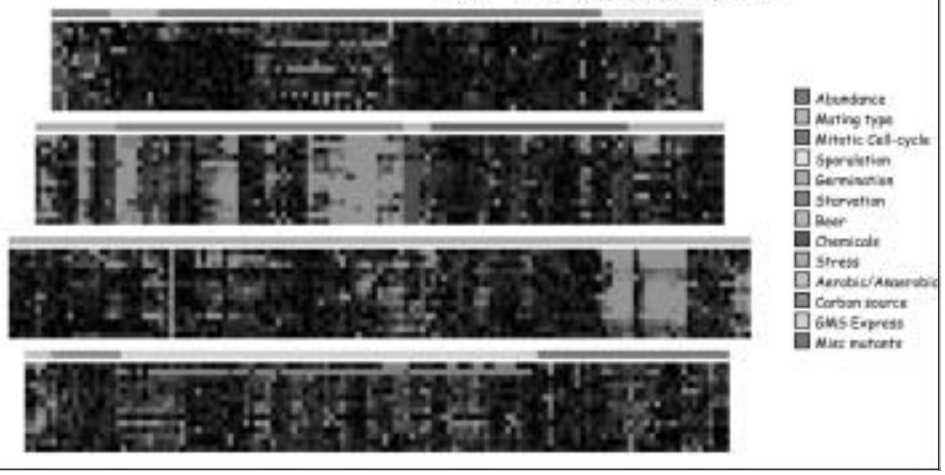
# Amino Acid and Purine Biosynthesis

Arg, Trp + misc



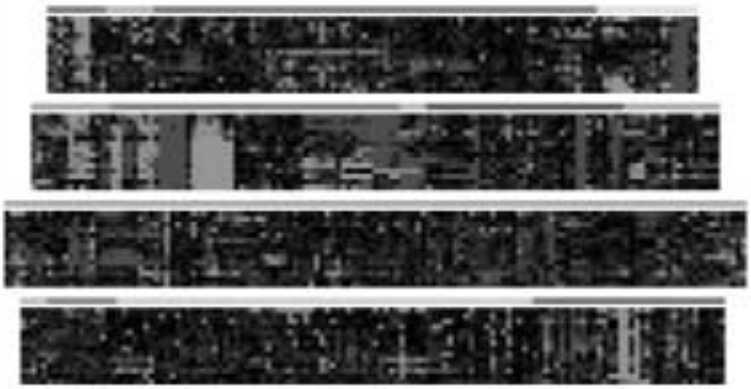

# Ergosterol Biosynthesis Cluster

| | |
|---|---|
| ERG5 | C-22 sterol desaturase |
| CYB5 | cytochrome b5 |
| ERG3 | C-5 sterol desaturase |
| ERG12 | mevalonate kinase |
| ERG10 | acetoacetyl CoA thiolase |
| ANB1 | translation initiation factor |
| ERG2 | C-8 sterol isomerase |
| MVD1 | mevalonate pyrophosphate decar |
| ERG13 | 3-hydroxy-3-methylglutaryl coenzyme A synthetase |
| ACS2 | acetyl-coenzyme A synthetase |
| ERG1 | squalene monooxygenase |
| ERG26 | sterol C-3 dehydrogenase |
| ERG6 | delta-24 sterol C-methyltransferase |
| ERG20 | farnesyl-pyrophosphate synthetase |



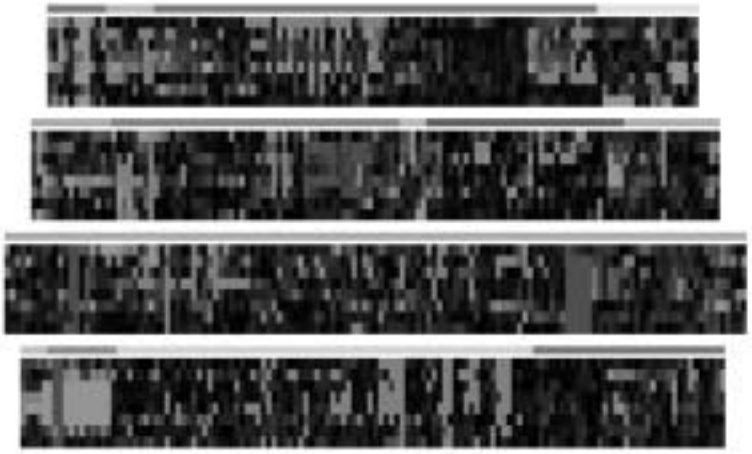
# Methionine Biosynthesis Cluster

| | | | |
|---|---|---|---|
| RAD59 | | MET10 | sulfite reductase subunit |
| MET32 | transcription factor | MET3 | sulfate adenylyltransferase |
| MET28 | transcriptional activator | MET14 | adenylylsulfate kinase |
| MET2 | homoserine O-acetyltransferase | MET17 | O-acetylhomoserine sulfhydrylase |
| SUL2 | sulfate permease | MET13 | methylenetetrahydrofolate reductase |
| MET1 | siroheme synthase | MET6 | homocysteine methyltransferase |
| MET16 | 3'-phosphoadenylylsulfate reductase | SER3 | 3-phosphoglycerate dehydrogenase |
| MET5 | sulfite reductase | MET22 | 3'2'5'-bisphosphate nucleotidase |

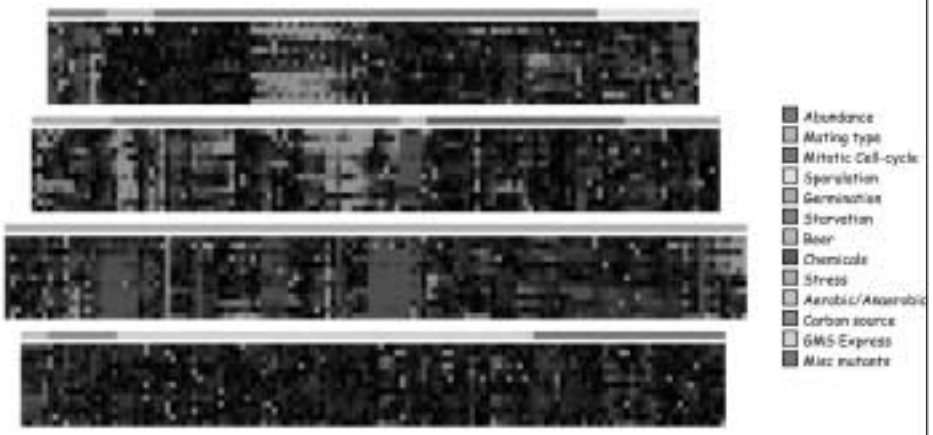



# Galactose Utilization Cluster

GAL1,2,3,7,10,80 plus
vessicular transport protein of unknown function

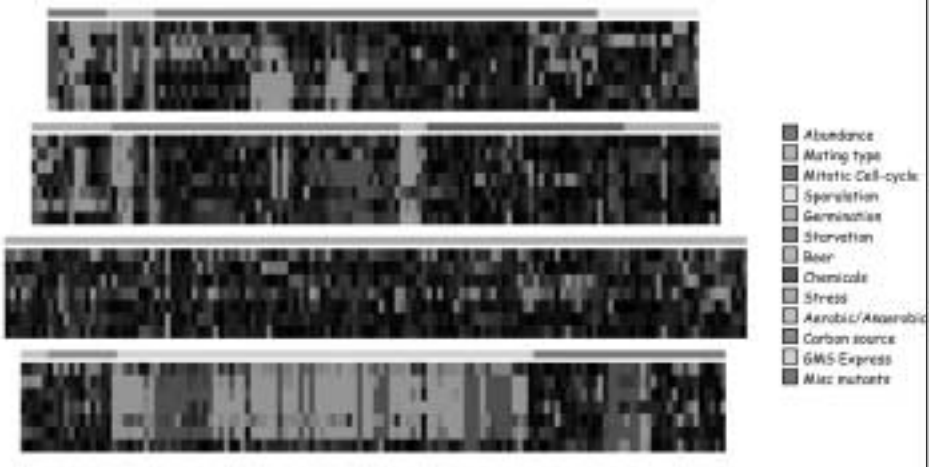


## Protein Folding Cluster



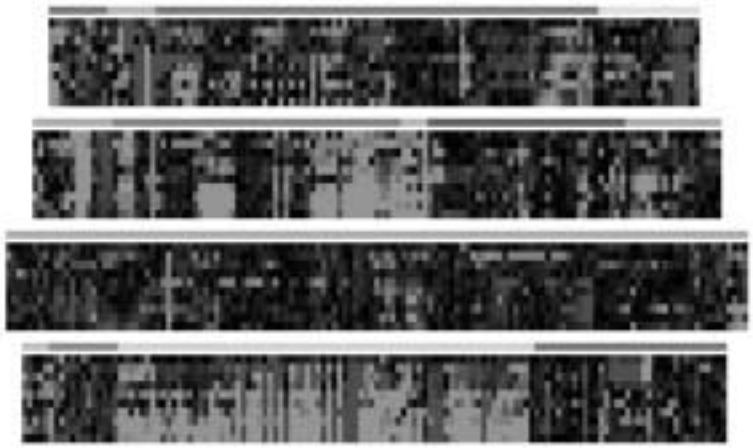

## Mating-type Alpha Cluster

| | |
|---|---|
| MATALPHA1 | alpha-specific gene activator |
| YLR040C | unknown |
| STE3 | a-factor receptor |
| MFALPHA1 | alpha factor |
| MFALPHA2 | alpha factor |
| SAG1 | alpha-agglutinin |
| LIF1 | unknown |



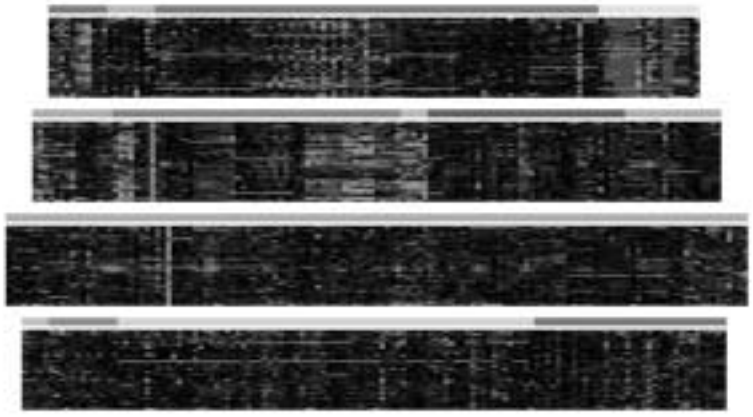
## Mating-type a cluster

| | |
|---|---|
| KAR4 | pheromone transcription |
| SST2 | alpha-factor desensitization |
| AGA1 | a-agglutinin anchor subunit |
| FUS1 | fusion; SH3 domain protein |
| MFA1 | a-factor precursor |
| AGA2 | a-agglutinin binding subunit |
| MFA2 | a-factor precursor |
| BAR1 | alpha-factor degradation |
| STE2 | alpha-factor receptor |
| STE6 | a-factor exporter |



Abundance
Mating type
Mitotic Cell-cycle
Sporulation
Germination
Starvation
Beer
Chemicals
Stress
Aerobic/Anaerobic
Carbon source
GMS Express
Misc mutants

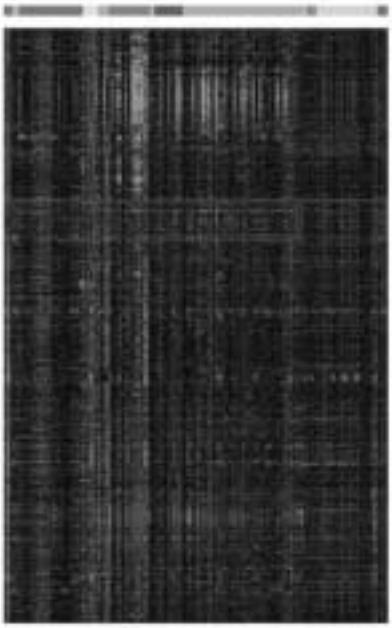
## Sporulation Cluster



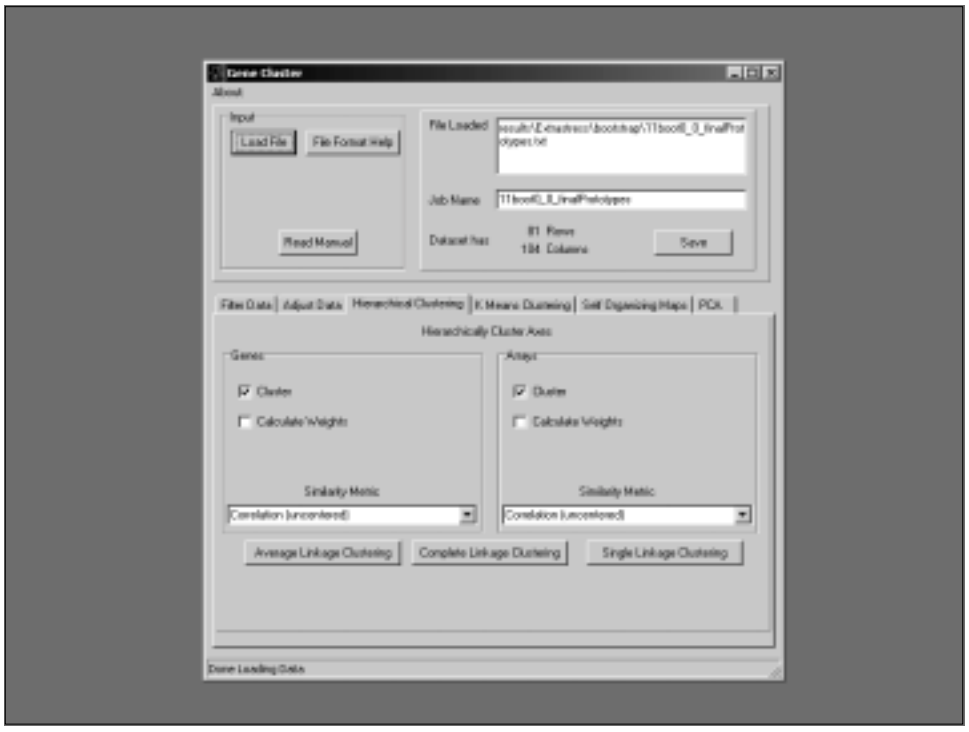
Abundance
Mating type
Mitotic Cell-cycle
Sporulation
Germination
Starvation
Beer
Chemicals
Stress
Aerobic/Anaerobic
Carbon source
GMS Express
Misc mutants

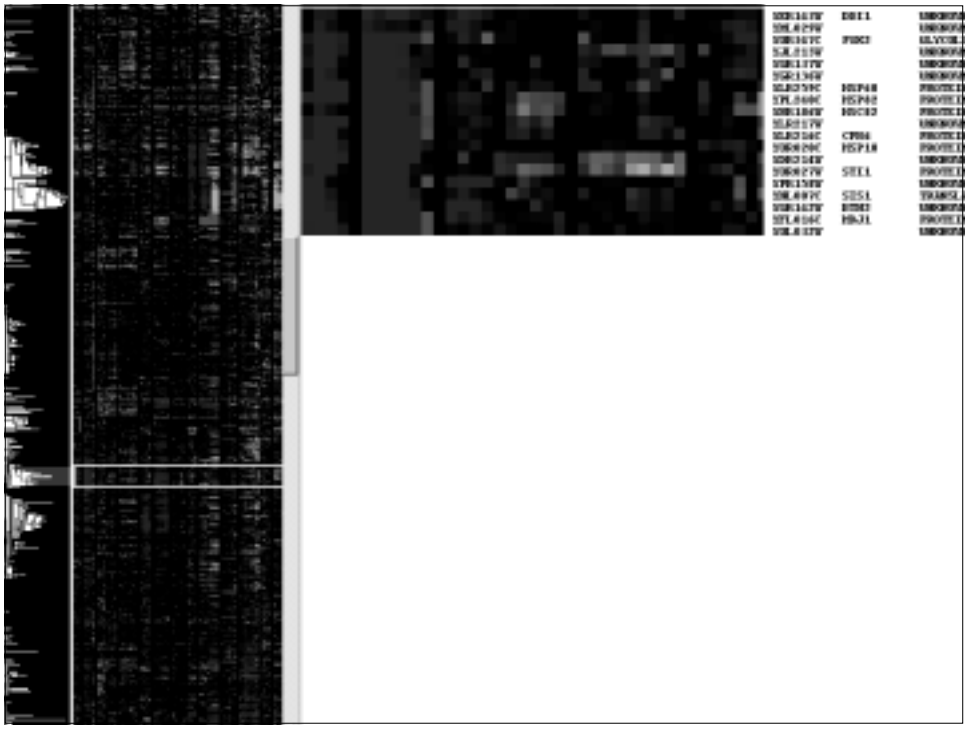
## Gene Expression Program of the yeast Saccharomyces cerevisiae

- Abundance
- Mating type
- Mitotic Cell-cycle
- Sporulation
- Germination
- Starvation
- Beer
- Chemicals
- Stress
- Aerobic/Anaerobic
- Carbon source
- GMS Express
- Misc mutants



**Genes of Unknown or Poorly Characterized Function**


Input
Load File
File Format Help
Read Manual
File Loaded
Job Name
Dataset has
81 Rows
184 Columns
Save
Filter Data
Adjust Data
Hierarchical Clustering
K-Means Clustering
Self Organizing Maps
PCA
Hierarchically Cluster Axes
Genes
Cluster
Calculate Weights
Similarity Metric
Correlation (uncentered)
Arrays
Cluster
Calculate Weights
Similarity Metric
Correlation (uncentered)
Average Linkage Clustering
Complete Linkage Clustering
Single Linkage Clustering
Done Loading Data



## Reading Genome Sequences

Great progress has been made in reading the protein coding information contained in genomes: identifying the location and structure protein coding genes, and of using the amino acid sequences to predict gene function and 3D structure

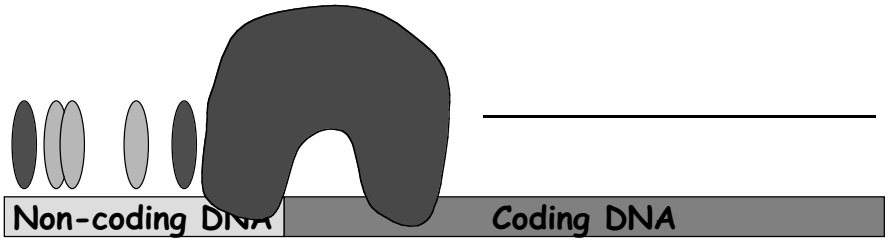
## Reading Genome Sequences

Comparably little progress has been made in reading the non-coding content of genomes, especially information that encodes when, where and under which conditions genes will be expressed.

## Reading Genome Sequences

The sequencing of the human genome was accompanied by the prediction of ~35,000 protein coding genes by two independent research teams, but essentially no predictions about the likely expression patterns of these genes.

## How is regulatory information written in genomes?



## Reading cis-Regulatory Code What do we need to know?

Output of system

Temporal, Spatial and Conditional gene expression patterns of all genes

## Reading cis-Regulatory Code What do we need to know?

Input to system

- cis-DNA sequence
- *in vitro* binding affinities of transcription factors
- *in vivo* binding distribution of transcription factors
- Evolutionarily conserved non-coding sequences

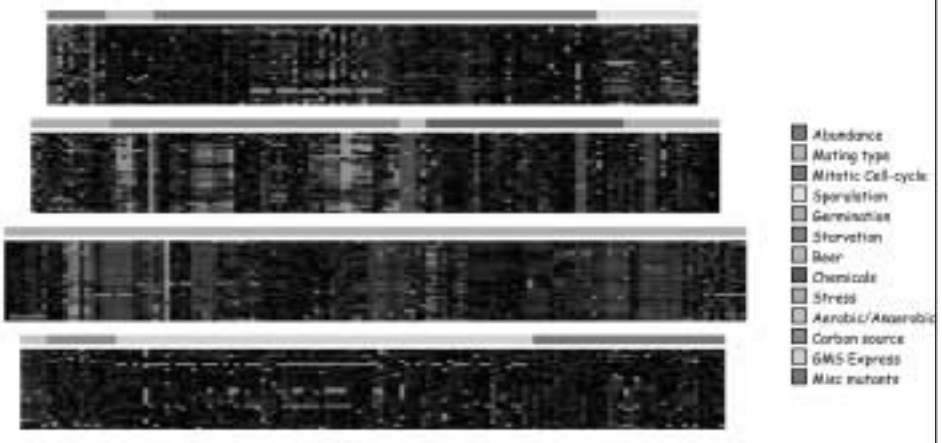
# Reading cis-Regulatory Code What do we need to know?

Input to system

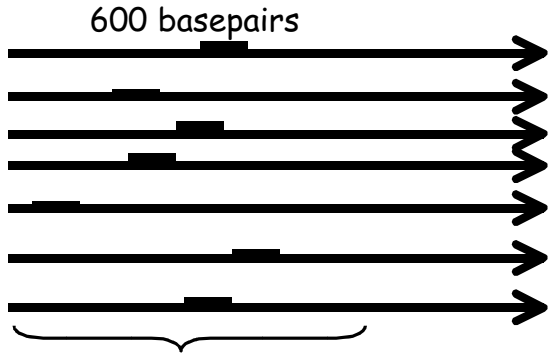
- cis-DNA sequence
- *in vitro* binding affinities of transcription factors
- *in vivo* binding distribution of transcription factors
- Evolutionarily conserved non-coding sequences

## Proteasome Cluster

Contains virtually all known components of proteasome (20s and 26s subunits)



# Cluster of co-expressed genes Pattern discovery in regulatory regions




Pattern over-represented in cluster

Possible examples of motif 1 in the training set

| Sequence name | strand | Start | Score | Site |
|---|---|---|---|---|
| YLR118C | w53 | 736 | 10.74 | AAATAAAAAT GTTAGCAAAA GAGAACGAAA |
| YDL126C | w53 | 660 | 17.33 | ATGTGAATTC GGTGGCAAAA AAGGACATAT |
| YER021W | c53 | 307 | 10.44 | TTTCTTTGAT GGCGGCAAAT ACTAATCTTA |
| YER021W | c53 | 705 | 17.07 | CTCTTTCTTG GGTGGCAAAT GTTTTTCGTC |
| YPR108W | c53 | 646 | 15.17 | ATCCTCAATG GGTGGCAAAC TTAACTTATT |
| YDR427W | w53 | 560 | 9.77 | TCACGCATCG TGTTGCAAAT AGCATATACA |
| YDR427W | w53 | 692 | 17.33 | GCTTTCATCC GGTGGCAAAA GTAAGAACAA |
| YOR157C | w53 | 643 | 14.30 | AACCCGGCCG AGTGGCAAAA TACCAAAAAG |
| YDL097C | w53 | 688 | 17.07 | TGGTTCCAAG GGTGGCAAAT GTTGGTAGAA |
| YOR117W | w53 | 637 | 17.33 | GCCAATATGT GGTGGCAAAA ATGAATTAAT |
| YML092C | w53 | 681 | 17.33 | AGCTTATTAA GGTGGCAAAA TTGTCCCCGG |
| YER012W | w53 | 659 | 17.33 | AAATCTTTAC GGTGGCAAAA AATAAAGAAA |
| YLR387C | w53 | 709 | 15.03 | GTCACAAGAC GGTGGCAACA CAGTTGTACA |
| YDR394W | w53 | 625 | 17.33 | CTGAAAGTAA GGTGGCAAAA TAGTAAATCT |
| YPR103W | w53 | 646 | 17.33 | ACAGCAAAAG GGTGGCAAAA CGAAGAATAG |
| YJL001W | w53 | 683 | 17.33 | AACGGAATCC GGTGGCAAAA AAGGGAAAAG |
| YFR050C | c53 | 689 | 17.07 | ACGTATTTAA GGTGGCAAAT TCTAAACCAA |
| YFR052W | c53 | 703 | 17.07 | AATTTCAATG GGTGGCAAAT AATGTGTAGA |
| YGL011C | w53 | 446 | 10.40 | TTAGGATTGT GGTGGTAAAA CTTGTAAAAT |
| YGL011C | c53 | 692 | 17.07 | CTATCCTTAC GGTGGCAAAT TGGTGTGTGC |
| YKL145W | c53 | 645 | 17.33 | TTAGTTTTCG GGTGGCAAAA TTCGGTAATC |
| YGL048C | w53 | 717 | 17.33 | ATGCGAGCGA GGTGGCAAAA GTTGTAATAT |
| YGR135W | w53 | 642 | 17.33 | TTACCCTGCA GGTGGCAAAA AATTTATAGA |
| YGR253C | c53 | 480 | 10.28 | CTTCAGCAGT AGTTGCAAAA GTGTCCTTTT |
| YGR253C | w53 | 697 | 17.33 | AACCAAGACC GGTGGCAAAA GAGTAACCTA |
| YBL041W | w53 | 714 | 17.33 | AAACTCAAAG GGTGGCAAAA ACAAGCGGTA |
| YFR004W | w53 | 677 | 17.07 | GGACCTCCAA GGTGGCAAAT TCACAAGAAA |
| YJL031C | c53 | 162 | 11.38 | GTGGCTTCTT GGTGGAAAAT GCAATCAATT |
| YJL031C | c53 | 675 | 13.79 | TGCTGTTTAA TGTGGCAAAT ATAAAAAATT |
| YMR004W | w53 | 169 | 12.36 | AAAATTATCA GGTCGCAAAT CAGATTTGCT |
| YMR004W | c53 | 182 | 10.21 | GATAAGTCTT TGTAGCAAAT CTGATTTGCG |
| YBR170C | w53 | 725 | 14.03 | CAGAACGAAG AGTGGCAAAT AATAGGACAG |
| YOR059C | w53 | 727 | 14.77 | ATAAAGAAAG GGTGGCAACT CTTTACTGCA |
| YKR014C | w53 | 168 | 10.78 | AATGTGGATT TTTGGCAAAT CAATGTTTCT |
| YJL053W | c53 | 465 | 10.08 | GACCCAGTTC TGTGGCCAAT CTTTTTCTTC |
| YPR107C | w53 | 557 | 15.17 | ATCCTCAATG GGTGGCAAAC TTAACTTATT |
| YGL047W | c53 | 113 | 11.64 | CGTTGATGAT GGTGGAAAAA GTTCCTGATT |
| YGL047W | c53 | 236 | 10.04 | ACATCGTGGA TGTTGCAAAA GATATAAACG |
| YGL047W | c53 | 610 | 17.33 | ATGCGAGCGA GGTGGCAAAA GTTGTAATAT |

**Proteasome Cluster -600 to -1**

**MEME (www.sdsc.edu/MEME)**

**RPN4** **Subunit of the regulatory particle of the proteasome**

| | |
|---|---|
| **Gene Name/Synonyms** | **RPN4; SON1; UFD5; GVM1; D2840; YDL020C** |

**At-a-Glance**

| | |
|---|---|
| **Cellular Role** | **Protein degradation** |
| **Biochemical Function** | **Proteasome subunit** |
| **Localization** | **Nuclear;19S regulatory particle of the proteasome** |
| **Mutant Phenotype** | **Null: viable** |

**mammalian homolog cannot be found in purified proteasomes**
**has two potential nuclear localization (NLS) sequences**
**not detectable in the purified proteasome preparation by**
**direct sequencing or by detection with antibodies**

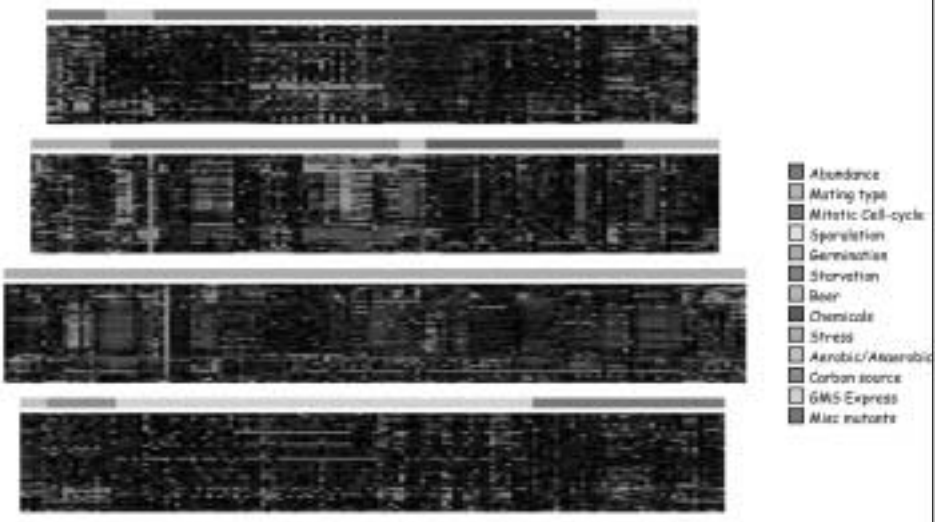
# GGTGGCAA is a binding site for RPN4

***FEBS Lett* 1999 Apr 30;450(1-2):27-34**

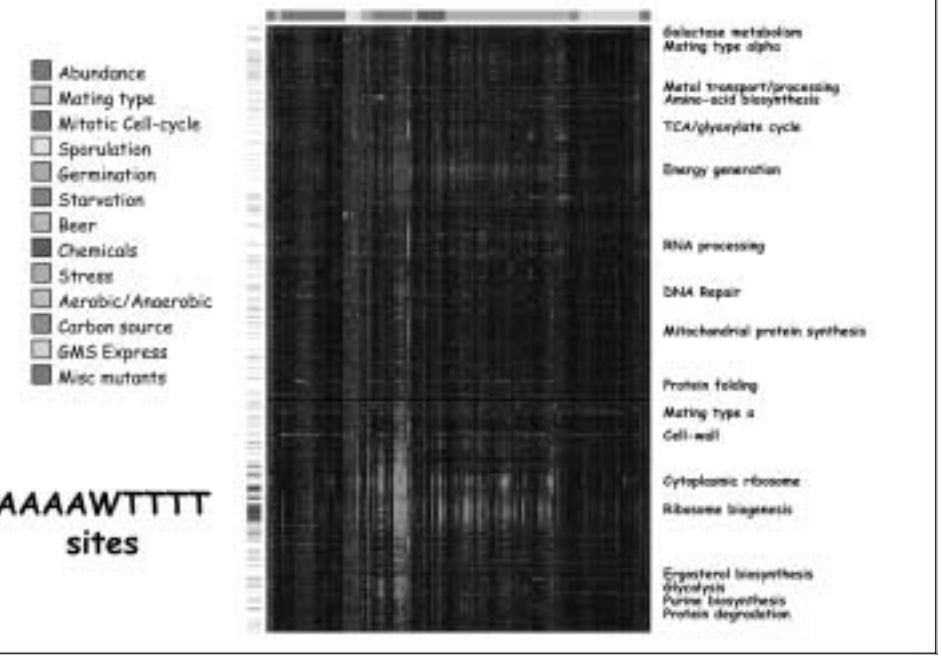
**Rpn4p acts as a transcription factor by binding to PACE, a nonamer box**
**found upstream of 26S proteasomal and other genes in yeast.**
Mannhaupt G, Schnall R, Karpov V, Vetter I, Feldmann H
Adolf-Butenandt-Institut der Ludwig-Maximilians-Universitat Munchen, Germany.

**We identified a new, unique upstream activating sequence**
(5'-***GGTGGCAAA***-3') in the promoters of 26 out of the 32 proteasomal yeast genes characterized to date, which we propose to call proteasome-associated control element. By using the one-hybrid method, we show that the factor binding to the proteasome-associated control element is Rpn4p, a protein containing a C2H2-

## Genes with Proteasome Control Element GGTGGCAAA




## Gene Expression Program of the yeast Saccharomyces cerevisiae



# Reading cis-Regulatory Code What do we need to know?

Input to system

- cis-DNA sequence
- *in vitro* binding affinities of transcription factors
- *in vivo* binding distribution of transcription factors
- Evolutionarily conserved non-coding sequences